UNIVERSITY OF CALIFORNIA
SANTA CRUZ

CMPS 5P – Introduction to Python in Programming

Instructor – Peter Alvaro

Sinjoni Mukhopadhyay

Assignment 6

Deadline: 03/06/2019 at 11.59 p.m.

Total Points: 100

Data Mining using Pandas

Instructions:

1. The aim of this assignment is to familiarize you with the Pandas tool that can be used to manipulate datasets and plot graphs to represent the dataset.
2. Students are allowed to use the Internet, if necessary, to figure out how to compute the diameter, perimeter, and area of a circle. However, they are NOT allowed to use the Internet to obtain code fragments. They need to write their program themselves.
3. Students should have all their code in a zip folder with name 'FirstLast_LastName.zip' containing the following ipython notebooks:
    a. Helper_Code.ipynb
    b. Travel-Reviews.ipynb
    c. Communities&crime.ipynb
4. Students should write code that is easily readable. Follow instructions provided carefully and make the appropriate changes. Failing to do this will result in losing points at the discretion of the grader.
5. For this assignment you have been provided with three ipython notebooks– one that contains some helper code and two that you need to fill in with your own code.
6. You need to run the first .ipynb file on Google Colab and attach the completed file with outputs.
7. For the other .ipynb files you need to modify it to generate expected answers for the questions asked.
8. Basic functions that you will need have been already imported for you in the .ipynb files. You will need to import the functions for PCA and the extra credit questions.

Helper Tutorials:

1. Installation guide for Pandas
https://pandas.pydata.org/pandas-docs/stable/install.html
2. Pandas Cookbook: gives an overview of Pandas functions
https://pandas.pydata.org/pandas-docs/stable/user_guide/cookbook.html
3. 10 minutes to pandas: short introduction to Pandas
https://pandas.pydata.org/pandas-docs/stable/getting_started/10min.html
https://towardsdatascience.com/23-great-pandas-codes-for-data-scientists-cca5ed9d8a38

4. Data Visualization
http://pandas.pydata.org/pandas-docs/stable/user_guide/visualization.html#visualization

Helper Code.ipynb

The Iris dataset:

In this assignment the example dataset we have used is the iris data by the UCI Machine Learning Repository. You can read more about the iris dataset and its attributes on:
https://archive.ics.uci.edu/ml/datasets/iris The helper ipython notebook is divided into three parts – data description, data preprocessing and data visualization:

A. Data Description: Before you begin processing your dataset, it is important that you know more about your datasets. Every dataset is divided into attributes and labels. Every label is defined by a set of attributes from the dataset. Here are a list of helpful functions for this section:

1. **.shape** – gives you the number of attributes and the total types of attributes
2. **.describe()** – provides you with a statistical summary of each attribute value in your dataset
3. **.hist()** – generates histograms to represent each attribute of the dataset in its raw form before data preprocessing.

B. Data Preprocessing:

Pre-processing allows us to manipulate the dataset to add/remove/modify values of the attributes. Here are a list of helpful functions for this section:

1. **.replace('a', nan)** – replaces occurance of a with NaN. Usually used to eliminate special characters from the dataset
2. **.dropna()** – drops the NaN values
3. **Normalizing/Scaling data** – A lot of times your dataset will contain features highly varying in magnitudes, units and range. For certain applications you need to scale your data to get a more uniform distribution in the attribute values. For this assignment we have used the minmax scaling function.
4. **.groupby() –** Usually used for Splitting the data into groups based on some criteria.

C. Data Visualization:

A basic plot of a dataframe can be made with the following lines of code:
df = df.cumsum() – returns cumulative sum over a DataFrame or Series axis
plt.figure();
df.plot.bar();

The basic code generates a line graph that represents the raw dataset. You can change the type of graphs by invoking the type function along with the plot function. Check out the example using the iris dataset that plots a bar graph representing the four features of the dataset.

Principal Components Analysis (PCA) for Data Visualization:

For a lot of machine learning applications it helps to be able to visualize your data. You can use PCA to reduce that 4 dimensional data into 2 or 3 dimensions so that you can plot and hopefully understand the data better.

For e.g. The original iris dataset has 4 columns (sepal length, sepal width, petal length, and petal width). In this section, the code projects the original data which is 4 dimensional into 2 dimensions. After the dimensionality reduction, there usually isn’t a particular meaning assigned to each principal component. The new components are just the two main dimensions of variation.

```
from sklearn.decomposition import PCA
pca = PCA(n_components=2)
principalComponents = pca.fit_transform(x)
#where x represents all attributes of the iris dataset
principalDf = pd.DataFrame(data=principalComponents,
columns=[‘principalcomponent1’, ‘principalcomponent2’])
```

Exercise 1:

Run the Helper_Code.ipynb and answer the following questions about the iris example dataset: [5 pts]

a. What are the attribute names of the the iris dataset? What are the three main classes of iris flower studied in the dataset? [2pts]
b. What do you think the .head() and .tail() functions do? [1pt]
c. Does the .groupby() function change the plot? If yes can you guess why it changes the plot? [2pts]

Exercise 2:

You will be performing all of your computations on the following two datasets:

1.Travel Reviews dataset:
https://archive.ics.uci.edu/ml/machine-learning-databases/00484/tripadvisor_review.csv

2. Communities and Crime dataset:
https://archive.ics.uci.edu/ml/machine-learning-databases/communities/communities.data

Write down python code, in the ipython notebooks provided to you, to find answers to the following questions:

Question 1 – Data description: [15pts]

1. Load your datasets from the links that has been provided to you above [3pts]
2. How many different types of attributes do each of the datasets contain? [3pts]
3. How many sample observations do each of the datasets contain for each attribute? [3pts]
4. What is the mean, max, min value and standard deviation of the attribute values for each of the attribute categories in both the datasets? [6pts]

Question 2 – Data Preprocessing: [20pts]

1. Display the values of the attributes in a dataframe format: [5pts]
   a. Category 1 and the User ID for the travel reviews dataset.
   b. The name of township attribute from the communities and crime dataset.
2. Do the datasets have any special characters as values? If it does then replace them with NaN and drop the values. [5pts]

3. Scale the following values using the sklearn min-max scaling function. Print out these values: [10pts]

   a. Last five category attributes of the travel reviews dataset.
   b. All the numerical attributes of the Communities and Crime dataset.

Question 3 – Data Visualization: [40pts]

1. Plot the raw dataset in the form of a box plot for both the datasets. [10pts]
2. Group the features and plot the new dataframe: [10pts]
   a. To reduce the total number of categories from 10 to 5 for the travel reviews dataset.
   b. By age of participants in the communities and crime dataset.
3. Plot the mean, minimum, maximum, standard deviation values of each numerical attribute in both the datasets. [20pts]

   Hint: -There are only 10 numerical attributes
   - The statistics need to be added to a dataframe before they can be plotted in a chart

Question 4 –PCA: [20pts]

Perform PCA on the dataset to reduce the number of observations and print the new dataset.

1. Reduce the number of categories from 10 to 3. [10pts]
2. Reduce the number of attributes from 128 to 40. [10pts]

Extra Credit: [20pts]

The document below explains how to perform PCA to speed up machine learning algorithms:

https://towardsdatascience.com/pca-using-python-scikit-learn-e653f8989e60

Read the document and use the guidelines provided in the document to perform the following functions on the communities and crime dataset that has been processed to remove special character values:

a. Split your dataset into training and test sets.
b. Transform your data using the techniques you have learnt in the assignment
   (standardize, scale, normalize).
c. Apply Logistic Regression to the transformed data.